PROCEEDINGS
OF THE
NATIONAL ACADEMY OF SCIENCES
Volume 16 October 15, 1930 Number 10

REFLECTION OF WAVES IN AN INHOMOGENEOUS ABSORBING MEDIUM

By Paul S. Epstein

Norman Bridge Laboratory of Physics, California Institute of Technology

Communicated July 21, 1930

1. *Introductory.*—The problem of wave propagation in a conducting and absorbing medium was approached from the point of view of geometrical optics in a preceding paper.[1] The purpose of it was to answer the question whether it is permissible to apply the ordinary rules for constructing the path of a ray to problems of radio-telegraphy, since certain layers of the atmosphere possess a considerable conductivity. The main result can be stated as follows: Theoretically speaking, the presence of conductivity has an effect on the geometry of the ray. Generalized equations of geometrical optics were obtained, showing that the rays are no longer normal to the wave surfaces. However, the difference is negligible numerically, unless the absorption is so large as to give an appreciable attenuation in a layer of the thickness of one wave-length.

This generalized discussion is instructive yet in another way, as the limitations of geometrical optics appear in it in a peculiarly drastic form. In an inhomogeneous medium, we have in general two waves: the refracted and the reflected one. The methods of geometrical optics permit us to trace directly only one of them which, for absorbing media, is always the refracted wave, as the phenomenon of total reflection does not exist in them. However, circumstances may occur in which the main part of the incident energy is contained in the reflected wave. In these cases, geometrical optics can give us only a very incomplete notion of the path in which the energy flows. It is, therefore, desirable to supplement that treatment by a discussion, what amount of reflection can be expected and what the conditions are in which the reflected rays may not be neglected. This is the purpose of the present paper: with regard to the problems of radio-telegraphy, the result is satisfactory in its simplicity. It can be shown that in a continuous medium, there is no appreciable reflection from a conducting inhomogeneous layer unless the conductivity is small and the conditions approximate those of total reflection. Ac-

cording to the preceding paper it is, then, permissible to neglect the conductivity altogether and to apply ordinary geometrical optics. This is exactly the procedure that has been followed in radio-telegraphic investigations, and our results give its complete justification.

The question of reflection in a continuously changing *transparent* medium was the topic of a considerable amount of discussion.[2] Under normal conditions the reflection is so small that some authors doubted its very existence. It is, therefore, interesting to investigate whether similar conditions prevail in the case of *conducting and absorbing* media. The following treatment is based on considering a law of change of the refractive index, (formula (8), section 3), which is not general, but of sufficient generality to give a good approximation to practically all cases that may occur in applications.[3] It has the advantage of mathematical elegance and rigor. In section 3, the rigorous general solution is found holding both for transparent and for conducting media. This solution is discussed, for a transparent medium, in section 4, and, for a conducting and absorbing medium, in section 5. It should be mentioned that we restrict our investigation to the ordinary form of the wave equation (Eq. (1) in section 2). If the waves are electromagnetic, this form applies directly to one state of polarization only (electric vector normal to the stratification of the medium). There is no question, however, that all the results will be qualitatively true also for the other state of polarization.

2. *Mathematical Formulation of the Problem.*—As mentioned above, the approximation represented by the methods of geometrical optics does not permit us to say anything about the reflected rays. We have, therefore, to fall back on the rigorous equation of wave motion

$$\Delta^2\psi + k^2\epsilon\psi = 0. \tag{1}$$

The notations are the same as (loc. cit.) ($k = 2\pi/\lambda$, ϵ the dielectric constant). For the purpose we have in mind, it will be sufficient to consider the case of a "*stratified medium*," i.e., the case when ϵ depends on one cartesian variable only $\epsilon = \epsilon(z)$. Moreover, we shall consider only cylindrical waves parallel to the direction y, so that ψ depends on x and z only. We can, then, make the substitution

$$\psi = e^{ikpx}\varphi(z), \tag{2}$$

obtaining for φ the plain differential equation

$$\frac{d^2\varphi}{dz^2} + k^2(\epsilon - p^2)\varphi = 0. \tag{3}$$

Of particular interest is the case, when the function $\epsilon(z)$ is such that it has a constant value ϵ_1 in a certain region, then changes, in a continuous

way, to another constant value ϵ_2. Without loss of generality, we can state this so: $\epsilon(z) = \epsilon_1$ for large negative values of z, and $\epsilon(z) = \epsilon_2$ for large positive values of z, with a layer of continuous transition in between.

Equation (3), being of the second order, has two particular integrals, P_1 and P_2, which we can choose in such a way that, *for large negative* z, they become asymptotically $\exp(ik\sqrt{\epsilon_1 - p^2}\, z)$ and $\exp(-ik\sqrt{\epsilon_1 - p^2}\, z)$. The corresponding solutions of (1) are

$$\left.\begin{aligned} [\exp(ikpx)P_1]_{z\to-\infty} &= \exp[ik(px + \sqrt{\epsilon_1 - p^2}\, z)], \\ [\exp(ikpx)P_2]_{z\to-\infty} &= \exp[ik(px - \sqrt{\epsilon_1 - p^2}\, z)]. \end{aligned}\right\} \quad (4)$$

They represent two plane waves going to and from the inhomogeneous layer in the lower medium.

On the other hand, we could select two particular integrals, of the same equation (3) in a different way. Let Q_1 and Q_2 be selected so that *for large positive* z they become $\exp(-ik\sqrt{\epsilon_2 - p^2}\, z)$ and $\exp(ik\sqrt{\epsilon_2 - p^2}\, z)$. Then

$$\left.\begin{aligned} [\exp(ikpx)Q_1]_{z\to\infty} &= \exp[ik(px - \sqrt{\epsilon_2 - p^2}\, z)], \\ [\exp(ikpx)Q_2]_{z\to\infty} &= \exp[ik(px + \sqrt{\epsilon_2 - p^2}\, z)]. \end{aligned}\right\} \quad (5)$$

They represent two plane waves in the upper medium. Any solution of equation (3) can be expressed as a linear combination of two particular integrals. We have, therefore, the relations

$$\left.\begin{aligned} Q_1 &= A_{11}P_1 + A_{12}P_2, \\ Q_2 &= A_{21}P_1 + A_{22}P_2, \end{aligned}\right\} \quad (6)$$

(with constant coefficients A), which are known to mathematicians as the "*monodromy group*" or the "*circuit relations*" of equation (3). Multiplying the second relation by $\exp(ikpx)$, we have

$$\exp(ikpx)Q_2 = A_{21}\exp(ikpx)P_1 + A_{22}\exp(ikpx)P_2.$$

This can be interpreted as a relation between the refracted (Q_2), the incident (P_1) and the reflected (P_2) waves. In other words, *this formula* represents the law of reflection, and the ratio A_{22}/A_{21} is the coefficient of reflection

$$R = A_{22}/A_{21}. \quad (7)$$

Mathematically speaking, the above formula gives the *analytical continuation* of function Q_2, represented by a power series, for instance, beyond the region for which this special representation is valid. The law of reflection gives us, therefore, a physical visualization of the rather abstract concept of analytical continuation. Originally, the theory

contained in the following sections was worked out by me because of this pedagogical value.[4]

3. *Solution for a Typical Case.*—By equations (6) and (7) the problem of reflection is reduced to that of finding the monodromy group of equation (3). This group is completely known only for one equation: the differential equation of the hypergeometric type. It is, therefore, advisable to approximate $\epsilon(z)$ by a function which will reduce (3) to the hypergeometric form and, at the same time, give a sufficient approximation to typical distributions of the dielectric constant, as they may occur in our physical applications. The most general function satisfying all our requirements is

$$\left.\begin{aligned} \epsilon(z) &= \epsilon_1 + e^{\zeta}[(\epsilon_2 - \epsilon_1)(e^{\zeta} + 1) + \epsilon_3]/(e^{\zeta} + 1)^2, \\ \zeta &= kz/s. \end{aligned}\right\} \tag{8}$$

When z is large and positive ($z \gg s\lambda/2\pi$), we have $\epsilon = \epsilon_2$, when it is large and negative, $\epsilon = \epsilon_1$. Roughly speaking, the parameter s represents the thickness of the inhomogeneous layer, measured in units $\lambda/2\pi$. Function (8) is, in fact, general enough for our purpose: If $-|\epsilon_2 - \epsilon_1| \leqslant \epsilon_3 \leqslant |\epsilon_2 - \epsilon_1|$, the change from ϵ_1 to ϵ_2 is monotonic. If this condition does not hold the dielectric constant first rises or falls to a maximum or minimum and then goes back to the final value. By a proper adjustment of the constant ϵ_3 the maximum or minimum can be given any value. At the same time, the thickness of the inhomogeneous layer can be varied by adjusting the parameter s, so that the law of change has a considerable adaptability.

In general the distribution is asymmetric with respect to $z = 0$. However, two special cases are of particular interest.

Case I: $\epsilon_3 = 0$,

$$\epsilon = \frac{\epsilon_2 + \epsilon_1}{2} + \frac{\epsilon_2 - \epsilon_1}{2} \operatorname{tgh} \frac{\zeta}{2}. \tag{9}$$

This is an antisymmetric inhomogeneous layer giving a monotomic transition from ϵ_1 to ϵ_2.

Case II: $\epsilon_2 = \epsilon_1$,

$$\epsilon = \epsilon_1 + \epsilon_3/4\cosh^2 \frac{\zeta}{2}. \tag{10}$$

For large negative ζ, the dielectric constant $\epsilon = \epsilon_1$, then it changes to a maximum or minimum $\epsilon = \epsilon_1 + \epsilon_3$, for $\zeta = 0$, and returns symmetrically to its original value $\epsilon = \epsilon_1$ when ζ becomes large and positive.

We transform equation (3) by choosing

$$u = \exp(\zeta) \tag{11}$$

as a new independent variable:

$$\frac{d^2\varphi}{du^2} + \frac{1}{u}\frac{d\varphi}{du} + s^2\left[\frac{\epsilon_1 - p^2}{u^2} + \frac{(\epsilon_2 - \epsilon_1)(u+1) + \epsilon_3}{u(1+u)^2}\right]\varphi = 0. \quad (12)$$

Moreover, we introduce the following abbreviations

$$\left.\begin{aligned} a &= -s\sqrt{p^2 - \epsilon_1} = is\sqrt{\epsilon_1 - p^2}, \\ b &= -s\sqrt{p^2 - \epsilon_2} = is\sqrt{\epsilon_2 - p^2}, \\ d^2 - d - s^2\epsilon_3 &= 0 \\ d &= \frac{1}{2} - \sqrt{\frac{1}{4} + s^2\epsilon_3} \end{aligned}\right\} \quad (13)$$

Putting $\varphi = (1+u)^d u^a f$, we obtain for f the equation

$$u(u+1)f'' + [(2d + 2a + 1)u + (2a + 1)]f' + (a + b + d)(a - b + d)f = 0, \quad (14)$$

showing that f is, in fact, a hypergeometric function. Hence, we obtain two solutions of (12) in ascending powers of u:

$$\left.\begin{aligned} P_1 &= u^a(1+u)^d F(a + b + d, a - b + d, 2a + 1, -u), \\ P_2 &= u^{-a}(1+u)^d F(-a + b + d, -a - b + d, -2a + 1, -u), \end{aligned}\right\} \quad (15)$$

and two solutions in descending powers:

$$\left.\begin{aligned} Q_1 &= u^{-b}\left(1 + \frac{1}{u}\right)^d F\left(a + b + d, -a + b + d, 2b + 1, -\frac{1}{u}\right), \\ Q_2 &= u^{b}\left(1 + \frac{1}{u}\right)^d F\left(a - b + d, -a - b + d, -2b + 1, -\frac{1}{u}\right). \end{aligned}\right\} \quad (15')$$

For large negative values of ζ and z, u becomes small and the approximations of P_1, P_2 are $u^a = \exp(ik\sqrt{\epsilon_1 - p^2}\, z)$, $u^{-a} = \exp(-ik\sqrt{\epsilon_1 - p^2}\, z)$, representing the incident and the reflected wave. Similarly, for large positive ζ and z, the approximation of Q_2 is $u^b = \exp(ik\sqrt{\epsilon_2 - p^2}\, z)$, representing the refracted wave. The part of the monodromy group (6), interesting to us, is given by a well-known Gaussian formula[5]

$$\left.\begin{aligned} Q_2 &= A(a - b, -a - b)P_1 + A(-a - b, a - b)P_2, \\ A(\alpha, \beta) &= \frac{\Gamma(\alpha + \beta + 1)\Gamma(\beta - \alpha)}{\Gamma(\alpha + d)\Gamma(\beta + d)} \end{aligned}\right\} \quad (16)$$

This leads to the coefficient of reflection

$$R = \frac{\Gamma(2a)\Gamma(-a - b + d)\Gamma(-a - b - d + 1)}{\Gamma(-2a)\Gamma(a - b + d)\Gamma(a - b - d + 1)}. \quad (17)$$

4. *Discussion for Non-absorbing Media.*—The coefficient R describes the reflected light as to its intensity and phase. R is, in general, a complex number and can be thrown into the form $R = |R| \exp(i\delta)$, where $|R|$ represents the amplitude of the reflected wave and δ its phase. In the case of non-absorbing materials ϵ is always real. The physical meaning of p is $p = n_1 \cos \varphi_1$, where φ_1 is the angle of incidence (counted from the glancing direction) and $n_1 = \sqrt{\epsilon_1}$. The parameter a becomes, according to (12), $a = isn_1 \sin \varphi_1$. The angle of incidence must, necessarily, be real so that a is always imaginary

$$a = ia' = isn_1 \sin \varphi. \tag{18}$$

The factor $\exp(ikpx)$ of equation (2) is the same everywhere, so that p can be interpreted also as $p = n_2 \cos \varphi_2$, where φ_2 is the angle of emergence of the refracted ray. The relation $n_1 \cos \varphi_1 = n_2 \cos \varphi_2$ is, of course, the law of refraction. The parameter b becomes $b = isn_2 \sin \varphi_2$, but φ_2 may be either real or imaginary, so that b, in its turn, may be imaginary or real. Finally, we see from (12) that d may be either real or have the complex form $d = 1/2 - id'$. Accordingly, we have to distinguish the following three cases:

(*A*) $a = ia'$, b real. Examining expression (17), we notice that the denominator is the complex conjugate of the numerator (both when d is real and when it has the form $d = 1/2 - id'$). This leads to

$$|R| = 1, \tag{19}$$

or to *total reflection*. The condition of total reflection is for a continuous inhomogeneous medium the same as for a discontinuous one: $p^2 \geqslant \epsilon_2$, $\cos \varphi_2 \geqslant 1$. This shows that we can have no total reflection when $\epsilon_1 = \epsilon_2$, as in the case of expression (9) or when ϵ_2 is complex. This confirms the result found (loc. cit.) by methods of geometrical optics.

(*B*) $a = ia'$, $b = ib'$, d real. Expression (17) can be transformed by means of a relation from the theory of gamma functions:

$$\Gamma(x)\Gamma(1 - x) = \pi/\sin \pi x. \tag{20}$$

Multiplying (17) by its complex conjugate and applying (20), we obtain

$$|R|^2 = \frac{\sin^2 \pi d + \sinh^2 \pi(a' - b')}{\sin^2 \pi d + \sinh^2 \pi(a' + b')}. \tag{21}$$

When s is small, i.e., when the inhomogeneous layer is thin, compared with $\lambda/2\pi$, we may replace the sin and sinh by their arguments. Moreover, a' and b' are proportional to s, while d becomes proportional to s^2. Neglecting fourth powers of s beside s^2, we obtain Fresnel's formula.

$$|R| = \left|\frac{a' - b'}{a' + b'}\right| = \left|\frac{n_1 \cos\varphi_1 - n_2 \cos\varphi_2}{n_1 \cos\varphi_1 + n_2 \cos\varphi_2}\right|.$$

On the other hand, when s is large, we can again neglect $\sin^2 \pi d$ and replace sinh by the exponential. We obtain $|R| = \exp(-2\pi b')$, for $a' > b'$, and $|R| = \exp(-2\pi a')$, for $a' < b'$.

In the particular case $\epsilon_1 = \epsilon_2$, we have $a' = b'$. The numerator is reduced to $\sin^2 \pi d$ and vanishes for integral values of d: this means that the reflected light exhibits the colors of thin or of thick plates.

(*C*) $a = ia'$, $b = ib'$, $d = {}^1/_2 - id'$. The same procedure as before leads to the formula:

$$|R|^2 = \frac{\cosh^2 \pi d' + \sinh^2 \pi(a' - b')}{\cosh^2 \pi d' + \sinh^2 \pi(a' + b')}. \qquad (22)$$

This case will not occur unless ϵ_3 is negative. It will apply to $\epsilon_3 < -|\epsilon_2 - \epsilon_1|$, s being sufficiently large, i.e., when the dielectric constant decreases to a minimum in the inhomogeneous layer. Geometrical optics (loc. cit., p. 44) would lead us to expect total reflection whenever ϵ decreases, in any place of the layer, below p^2. We see, however, from formula (22) that total reflection will not take place as long as ϵ_2 remains larger than p^2. The physical explanation of this fact is obvious: Even if we have total reflection ($\epsilon_2 < p^2$), the less dense medium is not optically vacant but contains a standing wave of the type $\exp(-k\sqrt{p^2 - \epsilon_2}z) \cdot \cos kp(x - ct)$, whose amplitude decreases exponentially with z. In the case of ϵ having a minimum, this standing wave reaches over into those parts of the field where ϵ, increasing again, becomes larger than p^2 and an advancing wave is again possible.

5. *Absorbing Medium.*—We shall adapt our discussion to the conditions prevalent in the case of radio waves. These waves originate in layers of the atmosphere which are practically free of conduction and absorption, so that we must regard ϵ as real and $a = ia'$ as purely imaginary. For ϵ_2 and ϵ_3, on the other hand, we shall make the general assumption

$$\epsilon_2 = \kappa_2 + i\Sigma_2, \quad \epsilon_3 = \kappa_3 + i\Sigma_3, \qquad (23)$$

where κ measures the refractive power and Σ the conductive power of the medium.

Correspondingly, we write for b and d

$$b = -b_2 + ib_1, \quad d = {}^1/_2 - d_2 - id_1 \qquad (24)$$

In the case of small conduction (i.e., when $\Sigma_2/|\epsilon_2 - p^2| \ll 1$ and $\Sigma_3/|\kappa_2| \ll 1$, so that we may neglect the squares of these small quantities) formulas (12) give us, for $\kappa_2 - p^2 > 0$,

$$b_1 = s\sqrt{\kappa_2 - p^2}, \quad b_2 = s\,\Sigma_2/2\sqrt{\kappa_2 - p^2}, \tag{25}$$

and for $\kappa_2 - p^2 < 0$,

$$b_1 = s\Sigma_2/2\sqrt{p^2 - \kappa_2}, \qquad b_2 = s\sqrt{p^2 - \kappa_2}. \tag{26}$$

In the same way, for $\kappa_3 + 1/4s^2 > 0$,

$$d_1 = s\Sigma_3/2\sqrt{\kappa_3 + 1/4s^2}, \quad d_2 = s\sqrt{\kappa_3 + 1/4s^2}, \tag{27}$$

and for $\kappa_3 + 1/4s^2 < 0$,

$$d_1 = s\sqrt{-\kappa_3 - 1/4s^2}, \qquad d_2 = s\Sigma_3/2\sqrt{-\kappa_3 - 1/4s^2}. \tag{28}$$

On the other hand, we obtain, for very large conduction ($\Sigma_2/|\kappa_2 - p^2| \gg 1$ or $\Sigma_3/|\kappa_3| \gg 1$)

$$\begin{cases} b_1 = b_2 = s\sqrt{\Sigma_2/2}, \\ d_1 = d_2 = s\sqrt{\Sigma_3/2}. \end{cases} \tag{29}$$

We have to evaluate the coefficient of reflection (17) for these cases. Such an evaluation cannot be given in a rigorous way for any value of s, but if s is sufficiently large to make the arguments of the gamma functions large compared with unity, it can be easily effected by means of the approximate formulas applying for large values of the argument x:

$$\begin{cases} \Gamma(x) = \sqrt{2\pi}\, x^{\,x-1/2} e^{-x} & \text{if } Re(x) > 0, \\ \Gamma(x) = \dfrac{\sqrt{2\pi}}{2 \sin \pi x}\, (1 - x)^{\,x-1/2} e^{-x+1}, & \text{if } Re(x) < 0. \end{cases} \tag{30}$$

Fortunately, the case in which these formulas apply is precisely that in which we are interested. We wish to learn something about the amount of reflection that is to be expected when dielectric constant and conductivity change continuously and slowly. We shall not give the general expression of the coefficient of reflection following from (17), because its discussion would become too cumbersome and would require too much space. It is sufficient for our purpose to discuss the two special cases mentioned in section 3: $\epsilon_3 = 0, d = 0$ (Case I) and $\epsilon_1 = \epsilon_2, a = b$ (Case II).

Case I.—The coefficient of reflection is reduced to

$$|\widehat{R}| = \frac{\Gamma(-a - b)\Gamma(-a - b + 1)}{\Gamma(a - b)\Gamma(a - b + 1)}. \tag{31}$$

The evaluation, by means of expressions (30), gives for the different cases:

(*Ia*) Σ_2 small, $\kappa_2 - p^2 > 0$; $b_2 \ll b_1$;

$$|R| = \left|\frac{b_1 + a'}{b_1 - a'}\right| 2b_2 e - (a' - b_1)\pi + |a' - b_1|\pi \tag{32}$$

(*Ib*) Σ_2 small, $\kappa_2 - p^2 < 0$; $b_1 \ll b_2$;

$$|R| = \exp\left(-4b_1 \operatorname{arctg} \frac{a'}{b_2}\right). \tag{33}$$

(*Ic*) Σ_2 very large; $b_1 \gg a'$, $b_2 = b_1$

$$|R| = \exp(-\pi a'). \tag{34}$$

Case II.

$$|R| = \frac{\sin \pi d}{\pi} \Gamma(-2a + d)\Gamma(-2a - d + 1). \tag{35}$$

(*IIa*) Σ_3 small, $\kappa + 1/4s^2 > 0$; $d_1 \ll d_2$

$$|R| = [\cos^2 \pi d_2 + \sinh^2 \pi d_1]^{1/2} \exp\left[-2\pi a' - d_1\left(\pi - 2 \operatorname{arctg} \frac{4a'}{2d_2 + 1}\right)\right] \tag{36}$$

In the case Σ_3 small, $\kappa + 1/4s^2 < 0$; $d_2 \ll d_1$,

we have to distinguish here two possibilities:

(*IIa'*) $d_1 < 2a'$

$$|R| = \left(\frac{2a' - d_1}{2a' + d_1}\right)^{d_2} \exp[(-2a' + d_1)\pi]. \tag{37}$$

(*IIb*) $d_1 > 2a'$

$$|R| = \left(\frac{d_1 - 2a'}{d_1 + 2a'}\right)^{d_2} \tag{38}$$

(*IIc*) Σ_3 very large; $d_1 \ll a'$, $d_2 = d_1$

$$|R| = \exp(-\pi a'). \tag{39}$$

Physical Discussion of the Results. Conclusion.—Let us begin the discussion with the case of small conductivity. Our formulas (*Ib*) and (*IIb*) refer to the case when refractive index and angle of incidence are such

that, neglecting conductivity altogether, we should have total reflection. We see that there is a continuous transition from the limiting case of nonconducting media. In fact, vanishing conductivity corresponds to $b_1 = 0$ and $d_1 = 0$, giving in both cases $|R| = 1$. If the conductivity is finite but small, we have, strictly speaking, not total but partial reflection. However, the coefficient of reflection is still very near to unity and all conditions are nearly the same as in the case of total reflection. On the other hand, if conditions are such that, neglecting conductivity, there would be a refracted and a partially reflected ray, formulas (32) and (37) apply, giving again results very close to those obtained in section 4 for transparent media. As the number $2a'$ or $2b_1$ appears in the exponent the amount of reflection is extremely small unless the angle of incidence is very close to the angle of total reflection ($b_1 = 0$). The case (*IIa*) has no analogy in a transparent medium but here, too, the amount of reflection is very small.

Going over to the case of high conductivity we find that the amount of reflection is always given by the simple expression $\exp(-a'\pi)$. This is in all practical cases a very small number as $a'\pi$ is large, provided that the inhomogeneous layer is a few times thicker than the wave-length, and that the incidence is not a glancing one. This result is particularly striking if we compare it with the conditions in a transparent medium. If the angle of incidence is larger, or slightly smaller, than the angel of total reflection, a non-conducting medium will give either total or considerable reflection. A slightly conducting upper medium will show practically the same behavior. With increase of conductivity the reflection decreases very rapidly (because of (33) and (26)), and when the conductivity is very large, the coefficient of reflection is extremely small. This seems paradoxical, at first sight, since we are accustomed to associate with metallic conduction a high reflectivity. However, a closer examination makes these facts look less surprising: in a continuous medium the reflected ray has its origin at all depths of the discontinuous layer, and the large absorption existing in this layer prevents its getting out of it with a considerable intensity.

Summarizing our discussion, we can say that the reflection is always very insignificant, except in the case when conductivity is small and where we have conditions very near to total reflection. This result indicates that in radiotelegraphy, if rays are reflected at all, their path can be computed neglecting conductivity, as if the medium were transparent.

[1] P. S. Epstein, *Proc. Nat. Acad. Sci.*, **16**, 37 (1930).

[2] The literature of this subject can be found in a paper by J. Wallot, *Ann. Physik*, **60**, 734 (1919).

[3] A special case of this theory was used by me ever since 1919 in my lectures on differential equations, as a physical illustration of the principle of analytical continua-

tion. For working out another special case, I am indebted to Dr. H. P. Robertson. I was told by the late Prof. E. Hilb of Würzburg, to whom I showed this work several years ago, that he had seen a paper by a French author proceeding on somewhat similar lines. However, I was unable to locate that paper.

[4] As an illustration of this theory, I usually demonstrate in my lectures the Wiener experiment (*Ann. Physik*, **49**, 105 (1893); Kohlrausch, *Praktische Physik*, p. 254, Leipzig (1921)) about the propagation of light in a medium composed of a layer of water over one of glycerine. It can be completely understood with the help of our theory of section 4 (with glancing incidence) and it is suitable for the determination of the relative index of refraction of water and glycerine.

[5] C. F. Gauss, *Werke*, **3**, p. 213, Göttingen, **1876**.

THE STRUCTURAL BASIS OF THE INTEGRATION OF BEHAVIOR

BY G. E. COGHILL

THE WISTAR INSTITUTE OF ANATOMY AND BIOLOGY, PHILADELPHIA

Read before the Academy April 28, 1930

The behavior of embryos has attracted the attention of biologists for a long time. William Harvey recorded important observations on the movements of the chick embryo. With the advent of the microscope Schwammerdam and Leeuwenhoek studied intensively the movements of snail embryos in the egg. Following their time there were only casual or isolated observations in this line until 1885 when Preyer published his work on the general physiology of the embryo, in which he gave large place to movements. Since Preyer, reports on the subject have been for the most part fragmentary; but not altogether so. Since 1920 Minkowski of Zürich has studied the movement of young human fetuses very extensively; and others, particularly Yanase, and Bolaffio and Artom, have made very important contributions to the knowledge of human fetal behavior. From all of this work, however, as it stands alone, no unifying principle or law has been deduced to resolve the knowledge of the development of behavior into an intelligible system. This is due, not to lack of acumen on the part of the observers, but to conditions inherent in the species studied by them; that is to say, the species have been high in the scale of evolution or morphological specialization, and have for the most part permitted of very limited time of observation. For some human fetuses the time of study has been limited to a very few minutes.

For the discovery of law or order in the development of behavior it is necessary to turn to a species which is relatively unspecialized morphologically and which can be studied continuously from the beginning of movement till the adult behavior pattern is established. In a complete historical or chronological panorama of the development of behavior of such a type the

fragmentary pictures from higher species fit appropriate places with a remarkable degree of conformity, and the conclusion seems justified that, from the lowest to the highest vertebrates, there is a common law according to which the individual attains its progressive adjustment to its environment. This law or pattern of development consists in the expansion of a primarily integrated total pattern of action within which partial patterns arise by individuation through restriction of both the field of motor action and the field of adequate stimulation. This law has been demonstrated in the origin of unconditioned reflexes, and it appears also to apply to the formation of conditioned reflexes, instincts and the so-called process of trial and error. The structural basis for this law is found in the growth of the nervous system; and in this also there appears to be a natural basis for the interpretation of attitude and motivation.

The salamander (Amblystoma punctatum), a morphologically generalized type of vertebrate, lends itself admirably to the study of the development of movement, for a particular individual of this genus can be observed continuously from the earliest muscular contraction till the definitive behavior pattern is established. Until some time after swimming begins this animal has no movable appendages. The earliest movement is a contraction of the most anterior muscles of the trunk. As the animal grows older the contraction becomes more extensive tailward till the whole trunk and tail are involved. The more extensive contractions always begin with the most anterior muscles and progress tailward as a perfectly integrated total reaction. A series of these flexures, alternating to one side and then the other, eventually effects aquatic locomotion. As appendages develop and acquire motility they move for a considerable period only as the trunk moves; and before local reflexes appear in the fore legs, or later in the hind legs, the alternating movement of the limbs in walking is established in coördination with the alternating movements of the trunk. Postural reactions of the trunk and limbs, also, are established before local reflexes of the legs appear. These reflexes gradually emerge from a total pattern of limb-and-trunk movements; and this occurs at about the time that antigravity action of the legs can first be observed. Posture and the walking gait as such are therefore primarily total reactions. Also the rotation of the leg in walking occurs first at the extreme phase of extension as a part of a total pattern of leg action.

The salamander orients itself in space also before local reflexes can be excited. This means that the primary relation of the vestibular system is with total reaction, not with local reflexes. Posture and the pattern of locomotion are, then, total systems. They are not built up synthetically by integration of reflexes. They arise by the expansion of a total pattern of action. Within them local reflexes arise by a process of analysis or individuation, and they appear in time primarily in relation to the

orientation of the animal to surfaces, rather than to space as such. They acquire varying degrees of independence according to the age, physiological condition and type of the animal concerned.

This order of development of behavior is explained by the order of growth of certain parts of the nervous system. The cephalocaudal expansion and progression of movement in the trunk have their origin in a series of motor neurones which develop progressively cephalocaudal and conduct in the same direction. From them the primary motor fibres go out to the trunk muscles. Branches of these fibres are the first motor fibres to the limbs, and at the beginning of movement of the limb these are the only perceptible motor fibres in the limb. As reflexes appear motor fibres from another region in the spinal cord join the primary motor fibres in the nerve roots. These secondary ventral root fibres presumably constitute the final common path for local reflexes, although the full demonstration of this is not yet established. The growth of the central and peripheral motor components of the nervous system, accordingly, involves the expansion of a unitary motor mechanism in which secondary systems emerge in subordination to the primary system; and into which essentially sensory systems (e.g., vestibular) grow and become consolidated or mechanized.

It is well known that, in the development of the reflex, there is a progressive restriction of the stimulogenous zone, or zone of adequate stimulus. In the early stage of development, stimulation at any sensitive point excites a total reaction. Within the total receptor field there emerge relatively large zones within which stimulation may excite a particular local reflex; and as development proceeds this zone of adequate stimulation becomes more and more localized. In the salamander, for example, early leg reflexes may be excited from an area on the trunk near the leg as well as by stimulation on the leg. In man (Minkowski) in early fetal stages the plantar reflex may be excited by stimulation on the dorsum of the foot as well as on the plantar surface, which is the definitive zone of adequate stimulation.

For an appreciable period before a particular receptor field acquires specificity in relation to an appropriate local reflex its stimulation inhibits the total reaction. Inhibition, accordingly, through stimulation of the exteroceptive field, begins as a total pattern. It is then in a field of total inhibition that the local reflex emerges. The reflex may, therefore, be regarded as a total behavior pattern which consists of two components, one overt or excitatory, the other covert or inhibitory. The essential anatomical basis for this is (1) in the mechanism of the total pattern of action, or primary motor system, and (2) in the mechanism of the local reflex, or secondary motor system; the mechanism of the total pattern being inhibited and that of the reflex excited. But since inhibition is not

a static condition but a mode of action, the mechanism of the total pattern must be regarded as participating in every local reflex.

In the light of these facts of structure and behavior, the fallacy of the traditional conception of the chain reflex, that one initial reflex stimulates the next in series, and this the following, etc., is apparent. In such an act as feeding, which may be regarded as a typical chain reflex, the process begins in the salamander with the elementary total reaction of swimming, which is followed in order by action of jaw, gill arches and gullet. But the motor neurones which excite these several phases of the act are in organic connection with the motor mechanism of the total reaction. In the first phase of the performance, the swimming reaction, the muscles of the jaws, gill arches and gullet, must be set, either by excitation or inhibition, for their responses in appropriate order. In this respect the situation is essentially like that of the earliest reflex that it is possible to elicit in the salamander: it can be excited only as the animal as a whole assumes a definite posture under the dominance of the mechanism of the total pattern. The particular acts of feeding, which may appear to be severally discrete, are actually only phases of a reaction which is fundamentally unitary. Chain reflexes are fundamentally total reactions under the dominance of the mechanism of the total pattern of action.

The conditioned reflex is conventionally regarded as differing essentially from the unconditioned reflex, but this is contradicted by evidence drawn from the development of behavior. Like unconditioned reflexes, the conditioned reaction emerges on the motor side from a field of general activity and on the sensory side the specific stimulus emerges from a general or wide zone of adequate stimulation. The latter phase of the development is obvious in a highly specialized form in conditioning to the metronome: the conditioning is first to the ticking in general and only later to a specific number of ticks. On the motor side of the conditioning, reaction is at first general, approximately a total reaction, at least of a postural nature, and only later does it become specific. Conditioning of reactions accordingly is accomplished by restriction (narrowing) of the zone of adequate stimulation and concomitant restriction in the field of action. The primary structural basis for this is in the mechanism of the total pattern.

The same interpretation can be applied also to learning by trial and error. The immediate response of an animal when it is placed in an utterly strange problem box or cage is general activity; and the stimulus to this action appears to be the situation as a whole. But within this general situation there eventually emerge particular features as relatively localized stimuli, and concomitantly out of the general field of action there emerges the particular act that is appropriate to the situation. This feature in the behavior of rats in the maze has been recognized by Professor Lashley.[1]

He says: "None of the studies of learning or retention of the mazes after cerebral lesions has given the slightest indication that the maze habit is made up of independent associational elements." And, elsewhere, "Many animals tend to explore the entire box at every trial until they become thoroughly adapted to the routine of training, and the initial improvement is largely due to this." In other words, the animal whose mechanism of the total pattern is particularly responsive to the general situation as a stimulus individuates the appropriate response out of the total pattern more readily than others do. Here again the mechanism of the total action pattern is the key to the process of learning.

Instinctive behavior is conventionally regarded as the result of compounding unconditioned reflexes essentially upon the principle of the so-called chain reflex. It involves the organism as a whole as opposed to the reflex, which is conventionally regarded as involving only a part of the organism. But, so far as is known, reflexes begin as total action patterns in response to a comparatively general field of stimulation and acquire their restricted function or specificity on both sensory and motor sides secondarily. Instincts typically are just such responses—total action patterns in response to relatively general situations. In proportion as they lose their pristine purity they acquire the nature of conditioned responses. Whether pure or modified, therefore, instinctive performances like reflexes have their origin in the mechanism of the total behavior pattern.

It has been explained above that the salamander acquires the ability to orient itself in space by the growth of vestibular neurones into the mechanism of the total pattern; but it should be noted that the vestibular system develops centrifugally, that is to say, the vestibular sense organs are the last elements of the system to maturate. The central relation of the neurones of the postural mechanism must therefore be determined without reference to the peripheral stimulation of the sense organs concerned. The same is true of the optic system. In fact, the major conduction paths of the diencephalon and mesencephalon are established in the salamander before there is any evidence of conduction paths entering these regions from the lower parts of the brain and before there is evidence of stimulation of the system through the olfactory, optic or other sense-organs. This cerebral mechanism, however, has at this pre-sensory period already established relations with the motor mechanism just as the vestibular component of the postural mechanism establishes its relation with the motor mechanism before the vestibular sense organs are developed. By this pre-sensory growth of the vestibular system the attitude which the individual will take toward space is predetermined; later when the vestibular sense organs are connected with this apparatus they merely furnish points of reference for orientation. The pre-sensory

growth of the cerebral mechanism may accordingly be regarded as determining what the attitude of the individual as a whole shall be toward the environment before the organism can take cognizance physiologically of its environment.

The determination of the primary attitude of the organism toward environment is intrinsic. When environment is brought to bear upon this intrinsic apparatus through sense organs, conditioning of behavior begins: the initiative of attitude is primarily within the organism. Not only does this statement hold for the initial response of the organism to sensory stimulation, but according to the mode of growth of the central nervous system, it probably holds also during later life; for after the sensory field has established connections with the motor system, and true sensory-motor processes are in play, neurones continue to develop in the brain outside the zones of simple sensory-motor activity. These neurones are, so to speak, supra-sensory or supra-associational elements. Their fibres always grow into the sensory-motor field of action. Their structure warrants the belief that they are functional conductors. That neurones may activate behavior before they are stimulated directly or indirectly from sensory fields in fishes and birds is well known. The hypothesis is therefore justified that these supra-sensory or supra-associational neurones in the growing nervous system are determiners of attitude or activating factors in behavior. The motivation of the organism as a whole must accordingly be, at least in part, intrinsic. The organism acts first on the environment and only later reacts to the environment. This principle is observed not only in behavior but in the growth of the nervous system. Any theory of motivation, therefore, which ascribes the activation of behavior wholly to the environment is inadequate. In so far as the correlation of structure and function in the development of behavior has been established there is justification for the belief in intrinsic sources of action on the part of the individual. The individual is more than reaction to its environment and it dominates its environment according as there is a growing component of supra-sensory and supra-associational elements in its nervous system. The structural basis for such a factor in behavior has been demonstrated in the salamander, though it is only in higher animals with well-developed cerebral cortex that this tendency comes to full expression.

In this discussion emphasis has been placed upon the organism as a whole and the mechanism of the total pattern. Although the conception of "the organism as a whole" may in a degree be mystical as used by some writers, the demonstration of a specific mechanism that at all times makes the normal individual a unit takes this conception out of the realm of mysticism or vitalism and places it on a scientific foundation. Much remains to be learned anatomically about the mechanism of the total

action pattern, which is the organ of the unity of the individual, but this much is known about it: it is a definite structure in lower vertebrates and it is a growing thing. Primarily it is strictly motor and its growth establishes the early behavior pattern without the intervention of sensory elements. Later elements that are categorically sensory (e.g., the vestibular system) grow into it and become inseparably mechanized with it to determine such a subtle action as the orientation of the individual as a whole in space. It is further known that neurones continue to grow into this mechanism from more and more remote fields until the cerebral cortex itself becomes consolidated with it in greater or less degree.

The mechanism of the total behavior pattern is, then, a growing thing. Its reality as such gives scientific grounds for the interpretation of the development of behavior, or learning, in general, as the expansion of a unitary system within which partial systems arise as dependencies under its jurisdiction: an interpretation that rationalizes the phenomena of behavior as treated under the various accepted categories.

[1] Lashley, *Brain Mechanisms and Intelligence*, University of Chicago Press, 1929. Pp. 134, 141.

LOCALIZATION OF FUNCTION IN THE NERVOUS SYSTEM

By C. Judson Herrick

Department of Anatomy, University of Chicago

Read before the Academy April 28, 1930

Some functions of the nervous system are clearly localized, typified by the classical conception of the reflex arc. About others we are not so sure, particularly the so-called higher or mental functions commonly ascribed to the cerebral cortex. Can their organs be arranged in mosaic patterns and mapped on the cortical surface like the charts of the old phrenology, or are they diffusely distributed so that they are impaired by loss of cortical substance in proportion to the amount of loss, irrespective of its location?

The unsolved problems of cortical localization may be clarified by an inquiry into the principles of localization in simpler and more primitive subcortical nervous tissue and the genesis of these primitive patterns. This theme may be approached from the standpoint of evolutionary development or of embryological development and the two lines of inquiry yield concordant results. Here we shall cite only the embryological evidence.

Of the numerous studies of development of behavior patterns, the most significant for the present purpose are those of Coghill on the salamander,

Amblystoma; for here we have precise records in statistically adequate number of the sequence of development of reflex patterns, together with the much more laborious histological examination of a series of specimens of tested physiological age. We know the order of appearance of reflex responses to stimulation, and also in considerable detail the structural changes that accompany the emergence of each new pattern of reflex and the exact location of these structures.

Dr. Coghill has pointed out that preceding the first neuromuscular response to stimulation there is an orderly development during which the unity of the organization is maintained by various non-nervous agencies, such as the mechanics of surface tensions, the chemical interplay of elementary hormones, differences in biolectric potential and physiological gradients. The neuromuscular apparatus develops within this unitary system and its earliest configuration is evidently shaped by it. The first nervous responses are integrated from the start; they are total reactions of the organism as a whole. There are no separate local reflexes; these come later.

The history of the emergence of some of the local reflexes from the total pattern has been presented by Coghill.[1] He has shown that the earliest movements of particular bodily members—limbs, jaws and so on—always occur in connection with total bodily movement. The capacity of a limb to respond reflexly to local stimulation independently of a generalized body movement is only gradually acquired. In this way the whole complement of individual local reflexes is slowly built up, these local patterns being successively emancipated from a more generalized type of response.

Parallel with this development of the behavior pattern, interesting changes are going on in the nervous system. The first neuromuscular reactions are executed through the mediation of definitely localized chains of neurons in synaptic relations so arranged that all sensory excitations are converged into the single final common path that discharges into the total body musculature, thus automatically integrating the behavior. This arrangement prevails—with some necessary complication of details—up to the perfection of the swimming reaction. Then follows, as an outgrowth from this integrated nervous system, the innervation of limbs and other apparatus of local movement. The same neurons which participate in the swimming movement send collateral branches outward into the growing limbs. The result is that the limb can move only when the body moves.

Up to the time of swimming the response at each stage of the development of the functional pattern is perfectly coördinated and integrated because it is performed by a definitely localized chain of nervous conductors so arranged that no other pattern of response is possible. It is

true that the response may be modified by the non-nervous factors of the total organization, for these are still operative; but no specific nervous machinery for modifiability of behavior has so far been demonstrated at this age.

Now as the local reflexes become emancipated from the total pattern and for each one of them a separate localized reflex arc is differentiated in the nervous system, the problem of the coördination of their activities and the maintenance of the unity of the organization takes on a new aspect. For no complication of separate and insulated reflex arcs, each of which is conceived as giving a one-to-one relation between stimulus and response, and no interconnection of such arcs by elaborate switchboard devices can conceivably yield the type of behavior which we actually find in higher vertebrates. The organism is reacting as a whole to its environment as a whole, and it is doing so in ways that cannot be formulated in terms of an algebraic sum or simple mechanical resultant of the interplay of the simple reflex responses to external stimulation. The behavior exhibits not only reaction, but also synthesis, initiative, originality—things that no combination of our hypothetical, simple, discrete reflexes can give. Even the simple response to a stimulus is not a reaction of a part of the body to a part of the environment. It is a response of the body as a whole (including as one factor the whole of its past experience) to the total situation in which the stimulus in question is set, and this response is fundamentally and primarily a unitary event, not an assemblage of separate sensorimotor components which are somehow and somewhere secondarily fabricated into an adaptive behavior pattern. The exponents of the Gestalt movement in psychology have made this very clear.

These facts are regarded as incompatible with the traditional dogmas of reflex physiology, with its precisely localized and well insulated reflex arcs and centers of reflex adjustment. The fundamental unitary character of this response, the preservation of its basic pattern amidst a kaleidoscopic shifting of the play of stimuli on the sensory field, together with its labile character and ready modifiability—for these and numerous other characteristics, the mechanisms of traditional reflexology seem hopelessly inadequate.

Yet reflex responses and sharply localized and well-insulated reflex arcs are unquestionably present. If it were not so, our present practice of clinical neurology and neurosurgery would be far less efficient than it is. These are the stock-in-trade of the neurologists. But perhaps this stock—as well as that of the psychologists—may be improved by the addition of some other wares without discarding any of those that they already have.

Let us now turn back and resume our examination of the differentiation of the nervous system of the growing salamander. At the stage when

local reflexes of limbs and jaws have become emancipated from their former subordination to mass movements of the body as a whole, histological examination shows that for each of these reflexes there is a definite nervous circuit. And it shows more than this. No one of these arcs is perfectly insulated from the others. There is no one-to-one relation between receptive organ and organ of response.

The neurons of the brain which are linked to form these conductors show remarkable complications. From the axons which form the main lines of direct conduction, innumerable collateral branches are given off. These may again branch repeatedly and run in diverse directions apparently haphazard. They rapidly increase in number, length and intricacy of ramification, and in the fields of their distribution other neurons are differentiated whose dendrites and ramified axons enter the entanglements of the collaterals first mentioned until there results a dense feltwork of finest nerve fibres which everywhere fills the interstices between the original pathways of the main lines of through reflex traffic. This is the "neuropil" within which the reflex arcs are embedded, and in the half-grown amphibian larva the specific and precisely localized reflex arcs are so completely immersed in this neuropil that it requires critical microscopic study to recognize them at all.

In the adult amphibian brain the diffuse neuropil and the nerve cells primarily related with it make up by far the larger part of the brain substance. The primary axons of the main lines of reflex conduction can usually be identified by their larger size, better myelination, or more direct course, but in many cases they can be followed through the diffuse and apparently random entanglements of the neuropil only by very careful study of especially favorable preparations. The separate and well-insulated reflex arc of neurological tradition has no real existence at any time in the history of this differentiation of the amphibian nervous system.

The various members of the body are innervated at first by collateral branches of the neurons of mass movement, and by the time the local reflexes are emancipated from these mass movements, their separate reflex arcs are all bound together through the diffuse neuropil so that no one of them can be activated without some influence being exerted upon the common matrix of all of them.

This diffuse and relatively equipotential neuropil, then, provides the anatomical substratum for the unity of the general behavior pattern, for the enduring individuality of the person, for the direct unanalyzed organismic quality of every separate sensorimotor response, for that plasticity and modifiability which are so characteristic of vertebrate behavior, i.e., for learning, and finally for the retention of more or less stable structural alterations of the reacting tissue resulting from particular patterns of response, i.e., for memory.[2]

The evidence for these sweeping general statements cannot be presented here, but attention may be directed to the fact that everything that is going on in the body is registered in this neuropil; into this tissue there is constant discharge of nervous impulses which overflow from every sensory path that enters the brain, and there is constant discharge from it to lower motor centers.

The neuropil of the adult salamander appears to be nearly homogeneous and physiologically equipotential, but not wholly so, for every part of it receives a preponderance of its nerve fibres from certain more or less specific sensory fields and has lines of preferential discharge toward certain motor fields. It is the parent tissue from which the specialized correlation centers, including those of cerebellum, thalamus, corpus striatum and cerebral cortex, of higher mammals have been differentiated; and the steps in the progress of this differentiation are preserved for our inspection in the organization of the series of brains from amphibian to human.[3]

The amphibian nervous system exhibits a definite pattern of localization embedded within a matrix of relatively unlocalized tissue serving the more general functions of the body as a whole. Turning to the human brain, we find this contrast accentuated. The number of individuated local reflexes is greatly increased and their nervous pathways and centers of adjustment are far more perfectly segregated and insulated. The result is that the human body has capacity for an immeasurably wider range of adjustments to every situation. There are more things that can be done about it, because there are more different kinds of response provided for in the stable bodily organization. Each of these local mechanisms is more highly specialized and is more definitely localized. This makes for greater efficiency in all stereotyped or automatized patterns of behavior.

Parallel with this specialization of the apparatus of local reflexes, there is a far greater elaboration of the integrating apparatus. Out of the diffuse neuropil there have emerged numerous special centers of adjustment, each of which in turn is more or less precisely localized, like the nuclei of the thalamus and the projection centers of the cortex. And there remains within these centers and between them an enormous mass of correlating tissue which preserves the primordial undifferentiated and labile character. This attains its maximum in the so-called associational mechanism of the cortex. This tissue is highly specialized, but its differentiation is in a direction opposite to that of the sharply localized apparatus of the stereotyped behavior patterns, for its prime function is the integration of these in patterns which fluctuate from moment to moment with every change in the content of the sensorimotor experience. This integration, moreover, is effected, not in terms of the present phase of

this experience merely, but in view of its own record of all previous experience.

Accordingly, the cortical apparatus of memory, apperception, imagination, invention and all constructive thinking cannot in the nature of the case be expected to be arranged in mosaic patterns whose boundaries can be mapped on the brain surface, like the charts of the ancient and modern phrenologists.

These are not disembodied functions, and the organs which perform them have form and location in space, but the forms which they assume are more like those of a continental system of telephone wires, or perhaps better, of wireless broadcasting, than the conventional pictures of the mosaic of cortical areas of, say, "word-memory" and "symbolic thinking." There are transmitters and receivers, whose location is fixed, but the apparatus connecting these may be as labile and as dynamic as the ether of interstellar space—and as imperfectly understood. The pattern of this apparatus may not be identical in two successive repetitions of the same externally evident sequence of stimulus and response, just as I may get telephone connection from Washington with an office in Philadelphia by way of Baltimore, or, in an emergency, by way of Harrisburg.

There is always some kind of localization of every cortical function. Some of these loci, like that of the electrically excitable "arm center," are tolerably accurately definable in terms of superficial cortical landmarks and they are tolerably constant from time to time. This is the classic mosaic localization. The apparatus of other sorts of cortical function does not submit to that sort of localization. By its very nature, it is widely diffused, inconstant, pliable; its configuration varies from moment to moment with a wide range of alternate or substitutional connections which give comparable, though not identical, results with every successive pulse of similar excitation. In response to a tap at my study door, I may say "come in" ten times in the course of a half-hour, but the chances are that the cortical process will not be identical the first and the tenth time.

Now, obviously, the word "localization" has been used in somewhat different senses in previous controversial literature, and even within the limits of this brief discussion. The anatomical localization of nervous functions of traditional neurology is adequate for clinical purposes up to a certain point. In the conventional reflex arc the terminal points are fixed, the peripheral pathways are fixed, and some, at least, of the central connecting tracts are known to be stable. The clinical diagnosis of a nervous lesion depends upon knowledge of the precise localization of these elements and upon their relative constancy.

But there are inconstant features even of the reflex. The scratch reflex of the dog is perhaps the best known example of a typical reflex,

thanks to the work of Sherrington. Here there is a wide area of skin within which stimulation of any point evokes the reflex and outside of which similar stimulation does not. This "receptive field" can be accurately mapped. The anatomical mechanism of this reflex pattern is so set up that adequate stimulation of any point within this field releases the reaction. In this simple case the locus of the receptor may be inconstant within this field. Then there are inconstant features within the adjusting center—the synaptic field of the spinal cord in this case. The typical response may be modified or inhibited by a variety of alterations of the physiological state of the center; but upon removal of the disturbing agencies the former pattern of performance returns.

De Silva[4] employs Henry Head's theory of "schematization" in an analysis of visual perception of movement in a way that may be helpful here also. The reflex pattern conforms to a schema which has a certain stability. Corresponding to this stable dynamic pattern there is an anatomical configuration of structural elements which, in the case of the scratch reflex, is known and localized. The apparatus of modifiability of the schema is not so well known, but the field within which this apparatus works can be delimited. The receptive field of the scratch reflex has such localization, and within this field the separate excitable points can also be located in each reaction. The synaptic field within the spinal cord, where inhbitions, etc., are effected, can also be delimited, but within this field we are quite ignorant of the configuration and precise location of the structural elements which determine what modification will follow some particular disturbing factor.

We have, then, two kinds of localization: (1) a known localization of stable structural elements whose functions also are known, and (2) a localization of fields within which various recurring patterns of performance, or schemata, are known to be fabricated and within which inhibition, modification, or conditioning of these patterns takes place. The first is the traditional functional localization of neurological practice as applied particularly in clinical neurology. The second is the sort of thing that is usually implicit in current diagrams of cortical localization—with very unsatisfactory results. The failures are due largely to our ignorance of cortical functions, and particularly to futile attempts to apply the criteria of localization of rigid structural elements to the second category of locations of patterns or schemata in fields.

It is possible to find out where the tissue is that yields these dynamic schemata and to delimit it; but these limits cannot be circumscribed on the surface of the brain in simple mosaic patterns. The manifestation of any schema at a particular time is always the function of a configuration of nervous elements which has location in space. But a very similar schema may at another time be exhibited by a different structural con-

figuration whose locus in space is by no means identical with the first. Our present technique is adequate to delimit the fields within which the schemata are operative, but it has not yet clarified the exact mechanism employed. Some of the simpler schemata are very stable, perhaps inherited; but others are evanescent. For the first group more or less of the apparatus can be identified as a stable anatomical pattern of fibre tracts and gray centers, but there is always a residuum whose structural basis is unknown.

We may, therefore, for convenience distinguish a localization of stable structural elements with the correlated localization of simple physiological processes from a localization of dynamic patterns or schemata within assignable fields. The latter may include fields of physiological gradients, fields of ionization or electric polarization, fields of differential chemical susceptibilities to hormones, etc., fields of organization of reflexes, fields of conditioning of these reflexes, fields where schemata of relation are fabricated (larger-smaller, lighter-darker, etc.) and in the human cortex fields of ideation, abstraction and the like.

The details of the mechanics of operation of these labile processes are as yet very obscure. Just what goes on in the associational tissue of the cerebral cortex is especially mysterious, but that this half-litre—more or less—of brain substance is the organ of integration of the personality and of control of behavior, including consciously purposive control, is as well-established empirically as anything in biology.[5] The unknown factors in this situation will probably be brought to light by an indirection, namely, by further study of the mechanics of nervous excitation and transmission and the fundamental properties of living substance in general.

[1] Coghill, G. E., *Anatomy and the Problem of Behavior*, Oxford Univ. Press, **1929.** Also *Proc. Nat. Acad. Sci.*, **16**, 637–643 (1930).

[2] Herrick, C. Judson, "Anatomical Patterns and Behavior Patterns," *Physiol. Zoöl.*, **2**, 439–448 (1929).

[3] Herrick, C. Judson, *Brains of Rats and Men*, Univ. of Chicago Press, **1926.**

[4] De Silva, H. R., "An Analysis of the Visual Perception of Movement," *Brit. J. Psychol.* (Gen. Sec.), **19**, 268–305(1929).

[5] Herrick, C. Judson, *The Thinking Machine*, Univ. of Chicago Press, **1929.**

GLYCOGEN METABOLISM*

BY WILDER D. BANCROFT AND GEORGE BANCROFT

DEPARTMENT OF CHEMISTRY, CORNELL UNIVERSITY

Communicated September 6, 1930

It seems to be agreed[1] that "glycogen is the fuel which the muscle uses for its activity....In a starving animal there is only a little glycogen, but there is always some. Without it an animal would not be able to move at all. Glycogen is used up during work, but it is restored again from sugar coming to the muscle-fibres through the blood. The blood gets the sugar either from the liver, where it is kept till needed, or from the intestines, where it is absorbed from the food during digestion. It seems probable that muscles cannot use directly any fuel other than glycogen or sugar."

"An element of muscular exercise, therefore, is conducted simply at the expense of carbohydrate: while, if the exercise be prolonged, other metabolic processes are stimulated, which presumably restore some of the carbohydrate at the expense of fat. Primarily, however, the fuel of the recovery process in the intact animal, as in the isolated muscle, is carbohydrate."

"Apparently, even when the stores of it are low, carbohydrate is used preferentially for the combustions following exercise. It is probable indeed that muscles really oxidize directly no fuel except sugar, and that other substances are oxidized only after transformation elsewhere, for example in the liver, into sugar. This, however, is a highly technical and hotly debated matter at present and justice cannot be done to it here."

"When a muscle is excited and fatigued, or sent into rigor by heat, chloroform, or other similar agencies, or allowed to undergo a prolonged process of anaerobic survival," there is a disappearance of glycogen and the appearance of an exactly equivalent amount of lactic acid, provided the process has been strictly anaerobic.[2]

"When a muscle contracts, a certain amount of lactic acid is liberated in it, in proportion to the amount of work done by the muscle. Very considerable amounts may be set free in the muscles of a man during hard work. An ounce of it may easily be formed when a man runs 100 yards as fast as he can. Now lactic acid is a pretty strong reagent, and there is no doubt today that fatigue is due mainly to its formation in the inside of the muscle-fibres. The function of oxygen is to remove the acid, once it has been formed."

"In a resting, isolated muscle left in oxygen, lactic acid does not appear: any originally there tends to disappear. In an isolated muscle stimulated in oxygen, lactic acid is set free but vanishes again during a subsequent

resting period. During this 'recovery process' so-called, CO_2 is produced equivalent in amount to the oxygen used, heat is liberated and glycogen is restored. The amount of glycogen gained is not equivalent to the lactic acid lost: a certain amount of one or the other is oxidized, presumably to provide the energy to drive the endothermic reaction:

$$\text{Sodium-lactate} + \text{protein} \longrightarrow \text{sodium-protein} + \text{glycogen}.$$

According to Meyerhof, about one molecule in four of the lactic acid removed in recovery is oxidized, the remaining three molecules reappearing as glycogen."

"Experiments on man have fully confirmed Meyerhof's observation on the isolated muscle, that carbohydrate only is oxidized in the recovery. How then are we to regard the recovery process? May we think of the

$$\text{Glycogen} \rightleftarrows \text{lactic acid}$$

mechanism as entirely distinct from the oxidative process in which energy is liberated to drive the reaction to the left? May we regard the breakdown and restoration as occurring in a separate system, analogous to a lead accumulator, having no connection with the oxidative reaction which provides energy to drive the dynamo—so to speak—by which the restoration is effected? Or should we consider some type of complex 'coupled' reaction, represented empirically by the formula:

$$5C_3H_6O_3 + 3O_2 \longrightarrow \frac{2}{n}(C_6H_{12}O_6)_n + 3CO_2 + 3H_2O.$$

It is impossible at present to say. The decision is one of the major problems of the science of carbohydrate metabolism: we can only indicate the problem now."

The questions, that A. V. Hill left unanswered in 1926, have been answered by Warburg and Meyerhof in one way, and we are answering them in the other way.

Levene and Meyer[3] showed that leucocytes and kidney tissue convert glucose into dextrorotatory lactic acid and form a mixture of the *dl*- and *d*-forms from methyl glyoxal under aseptic conditions. From this they concluded that glucose must break down into lactic acid by way of methyl glyoxal. Meyerhof[4] criticizes the views of Levene and Meyer, saying: "They seem to assume that a reversible equilibrium exists between sugar and lactic acid, so that the reaction could go spontaneously either in one or the other direction. This is not the case, however, as will be shown in detail later. The cleavage of sugar into lactic acid is a spontaneous process going to completion. On the other hand the synthesis of sugar from lactic acid requires a supply of energy furnished in the isolated muscle exclusively by oxidation of part or the corresponding amount of sugar. It can probably be provided also in the other organs only by oxidation."

In the same monograph, p. 55, Meyerhof says: "In the oxidative phase one molecule of sugar or the corresponding amount of lactic acid is burned. The rest of the acid is reconverted with phosphate to the ester (glucose phosphoric ester) and again becomes glycogen. Here we have a coupled reaction similar to the alcoholic fermentation."

From these and other paragraphs it is evident that Meyerhof does not believe in a reversible equilibrium between glycogen and lactic acid and that he believes that the conversion of lactic acid into glycogen is coupled with, and dependent on, the oxidation of lactic acid (or glucose).

In 1927 A. V. Hill[5] had apparently adopted Meyerhof's point of view. "In the recharging process in muscle,... a certain amount of fuel is burnt to supply the energy for the recharge. This fuel seems to be either glycogen or sugar, which disappears gradually from the muscle during a long bout of work—unless it be replaced from the blood. During the recovery process the lactic acid vanishes, but is found over again in its original form as glycogen. We may summarize the matter as follows:

Contraction

Glycogen [or glucose]⟶lactic acid + energy.

Recovery

[Lactic acid + oxygen⟶carbonic acid and water + energy]
Lactic acid⟶glycogen − energy.
Glycogen + oxygen⟶carbonic acid and water + energy."

Hill[6] finds that, under normal conditions, about one part of lactic acid is burned to four converted to glycogen. "It is possible by a direct analysis of the lactic acid in the blood to ascertain the concentration in the muscles and other soft tissues of the body at two moments: (1) about ten minutes after exercise, and (2) towards the end of recovery. By such direct analyses, and assuming that about fifty per cent of the body weight is in intimate contact with the blood, it is possible to determine directly—if somewhat roughly—the amount of lactic acid which has disappeared in the recovery process. This may be compared, as in the case of the lactic acid calculated from the CO_2 retention, with the oxygen used in that removal, and it is found, as in the frog's isolated muscle, that the acid is not oxidized; only about one-fifth of the oxygen is used which would be required were all the lactic acid broken down to CO_2 and water. The same ratio for the 'efficiency of recovery' is attained in man as in the isolated muscle."

"If the removal of the lactic acid were due simply to oxidation, the total heat in the whole process would be 3836 calories, namely, the heat of oxidation of one gram of glucose. This is about 4.5 times as great as the heat naturally set free, so that only 0.224 gram of glycogen is actually

oxidized in the complete cycle by which one gram of lactic acid is produced and then restored again as glycogen. This result agrees well with actual chemical determinations....We may calculate that an 'oxygen debt' of one liter implies the presence in the body of six grams of lactic acid. This value has been used in the calculations given in earlier chapters. Its use assumes that we may apply to human muscles the results obtained in the isolated muscles of the frogs. There is much evidence that this assumption is justifiable: probably the same chemical process is involved in all cases."

The ratio of lactic acid burned to lactic acid converted into glycogen may, in special cases, become 1:6. With cancerous tissue, infants and the brains of rabbits, the ratio changes toward the value 1:0, nearly all the lactic acid being burned or carried off in the blood, and little or none converted into glycogen.

It is not true, as Meyerhof postulates, that the conversion of glucose or glycogen to lactic acid is a spontaneous reaction going to completion. So far as anybody now knows, a solution of glucose in pure water would remain unchanged for an indefinite period. One must postulate either an enzyme or some other catalytic agent. Meyerhof rejects the idea of an enzyme, apparently forgetting that the lactic acid obtained is optically active. No method of change, that we now know and that does not involve enzymes, will give an optically active acid. If glucose or glycogen passes through the stage of pyruvic acid or of methyl glyoxal, both optically inactive substances, nothing but an enzyme can give optically active lactic acid. Such enzymes are known. Professor Peterson of the University of Wisconsin can prepare dextrorotatory or laevorotatory lactic acid from glucose by enzyme action. We are indebted to his courtesy for samples thus made. If Meyerhof postulates an enzyme in order to get an optically active lactic acid, he must postulate theoretical reversibility and then there is nothing left of his theory.

The conversion of lactic acid into glycogen cannot be a coupled or induced reaction in the sense that physical chemists use the term, because there is neither oxidation nor reduction involved in the reaction. Also, no such conversion ratios are known in any coupled reaction, so far as our knowledge goes.

If the two reactions are really coupled, it should be possible to convert lactic acid into glycogen by means of any substance that would oxidize lactic acid and would not oxidize glycogen. A number of attempts were made to do this, all unsuccessfully.

We believe that it is more rational to postulate no connection between the oxidation of lactic acid and the conversion of lactic acid into glycogen or glucose—through a phosphate if necessary—beyond the chronological one that both reactions often occur simultaneously.

If we are dealing with a reversible equilibrium between glucose or glycogen and lactic acid, brought about by enzymes, there is no difficulty about an optically active lactic acid, and the ratio between lactic acid burned and lactic acid converted may obviously have any value, depending on the conditions of the experiment.

On the other hand we have to account for the reaction running one way on stimulation under anaerobic conditions and the other way on recovery under aerobic conditions. It was the apparent impossibility of doing this that drove Meyerhof to the other explanation. The true answer was given or implied by Bayliss.[7] "It might be thought that a synthesis of a small degree could not be of much practical importance. This would be an error, as the following considerations will show. Let us take the case of amylase where the hydrolysis progresses almost to completion, and let us suppose that no more than one per cent of starch is formed when the enzyme acts on maltose or dextrin. Since the product is an insoluble body, the equilibrium will exist only for a moment, so that more starch will be formed in order to replace that thrown out of the system by precipitation. As the rate of this reaction is slow, as shown above, the amount of starch per unit time will not be great, although by no means negligible. The process, it will be noted, is analogous to that of the precipitation of chloride as silver salt. It is most likely, as Croft Hill[8] points out, that the storage of starch in the plant and that of glycogen in the animal are to be explained on these lines."

Strong adsorption of a soluble or peptized substance is equivalent to sparing solubility. Przylecki and Wofcik[9] have shown that protein has an extraordinarily high adsorptive capacity for glycogen, and that this adsorption may amount to ninety per cent with ten per cent glycogen using amounts of protein corresponding to those in the liver. With 30 cc. of a 1% glycogen solution and 30 grams protein, 99% of the glycogen is removed from the solution by adsorption. This adsorption of glycogen is a reversible reaction, for the glycogen can be removed from the protein by adding enough water. Alcohol and other narcotics also displace the glycogen, which accounts for the effect of chloroform in producing lactic acid. Przylecki worked chiefly with the enzyme amylase which converts glycogen to glucose, but the principle is the same.

In order to account for their results on the rate of formation of lactic acid, Fletcher and Hopkins[10] suggest the possibility of an equilibrium between free and adsorbed glycogen. "Conceivably the store of precursor in the muscle is partly in insoluble form, partly in solution; the concentration might be kept constant by replacement from the insoluble store, and, for a period, conditions would exist for a linear rate of change. The explanation is probably less simple than this, and it is interesting in any case to observe how in the quiescent unstimulated muscle the survival

processes which lead to lactic acid production are so controlled as to lose the exponential character of an isolated chemical reaction."

If we assume that a stimulated muscle adsorbs less glycogen, that will necessarily lead to a formation of lactic acid which is really a measure of the change in adsorptive capacity of the muscle and not a measure of the work done by the muscle, unless these two happen to be proportional, which we do not know to be the case. As the muscle returns to its normal state, the adsorption of glycogen increases and consequently there is a conversion of lactic acid into glycogen. The return to the normal state occurs more rapidly in the presence than in the absence of oxygen. There are no data to show whether there is any change in the effective concentration of the enzyme at any time.

In the case of cancerous tissue, Meyerhof assumes that lactic acid is converted into glycogen under aerobic conditions at the same rate as in normal tissue; but that, in some unspecified way, the glycogen is converted back into lactic acid as fast as formed. This is absurd. Meyerhof's explanation is certainly wrong, even if ours should not be right. Our assumption is that the resting cancerous tissue does not adsorb glycogen strongly and that therefore little or no glycogen is formed. It is probable that this assumption can be tested directly by adsorption experiments on cancerous and on normal tissue; but we are not equipped to do this. Since rabbit's brains[11] act similarly to cancerous tissue in this respect, the inability to convert lactic acid into glycogen must be an accompaniment of cancer and not a cause of it.

The general conclusions of this paper are as follows:

1. Warburg, Meyerhof and A. V. Hill assume that the oxidation of lactic acid and the conversion of lactic acid into glycogen are coupled reactions. We assume that they are independent in fact though not in time.

2. Meyerhof's theory requires the formation of an optically neutral lactic acid; and the lactic acid of the muscles is dextrorotatory.

3. If the two reactions are coupled, it should be possible to synthesize glycogen by the oxidation of lactic acid in a test tube. No one has ever done this and our attempts were unsuccessful.

4. We postulate a reversible equilibrium between glycogen or glucose and lactic acid, maintained by enzymes. This permits of glucose being converted into dextrorotatory or laevorotatory acid, depending on the nature of the enzyme, and requires the formation of optically neutral lactic acid when the reaction is caused or catalyzed by alkali, as in the experiments by Evans. These predictions have been fulfilled experimentally.

5. The apparent displacement of equilibrium is due to changes in the adsorption of glycogen by protein. We postulate that a stimulated

muscle adsorbs less glycogen than an unstimulated one. There is, therefore, a formation of lactic acid from glycogen when a muscle is stimulated and a formation of glycogen from lactic acid when the muscle returns to its normal state.

6. The power of the muscle to adsorb glycogen is recovered more rapidly in presence of oxygen. It is not known whether there is any change in the concentration or effectiveness of the enzyme under these conditions.

7. It has been shown experimentally that alcohol, chloroform and other narcotics displace glycogen from protein and that they cause formation of lactic acid, as they should.

8. The behavior of cancerous tissue can be accounted for on the assumption that it always shows a low adsorption of glycogen. There is no independent confirmation of this as yet. Meyerhof's assumption seems a contradiction in terms, that cancerous tissue tends to regenerate glycogen from lactic acid at the same rate as normal muscles; but that the rate of formation of lactic acid from glycogen is increased enormously in some obscure way.

9. Since rabbit's brains also have little or no power to convert lactic acid into glycogen, this inability of cancerous tissues to store up glycogen must be an accompaniment but not a cause of cancer.

10. The explanation that we have given harmonizes with the general views of Bayliss, Fletcher and Hopkins, and Levene and Meyer.

Our thanks are given to Dr. Ellice McDonald of the Cancer Research Laboratory of the University of Pennsylvania. He called our attention to the problem, made it possible for us to carry out the work and rendered invaluable assistance in furnishing us literature references.

* This work is part of the programme now being carried out at Cornell University under a grant from the Heckscher Fund for the Advancement of Research, established by August Heckscher at Cornell University. The expense has been met by a grant from the Cancer Research Laboratory of the University of Pennsylvania.

[1] A. V. Hill, *Living Machinery*, 1927, p. 62; *Muscular Activity*, 1926, p. 104; *Muscular Movement in Man*, 1927, p. 38.

[2] A. V. Hill, *Muscular Activity*, 1926, p. 58; *Living Machinery*, 1927, p. 63; *Muscular Activity*, 1926, pp. 65, 68.

[3] Levene and Meyer, *J. Biol. Chem.*, **12**, 265; **14**, 551 (1913).

[4] *Chemical Dynamics of Life Phenomena*, 1924, p. 59.

[5] A. V. Hill, *Living Machinery*, 1927, p. 67.

[6] A. V. Hill, *Muscular Activity*, 1926, p. 103; *Muscular Movement in Man*, 1927, p. 78.

[7] Bayliss, *The Nature of Enzyme Action*, 1925, p. 56.

[8] Croft Hill, *J. Chem. Soc.*, **73**, 634 (1898).

[9] Przylecki and Wofcik, *Biochem. J.*, **22**, 1302 (1928).

[10] Fletcher and Hopkins, *J. Physiol.*, **35**, 275 (1907).

[11] Holmes and Holmes, *Biochem. J.*, **20**, 1196 (1926).

NOTE ON THE NATURE OF COSMIC RAYS

BY PAUL S. EPSTEIN

CALIFORNIA INSTITUTE OF TECHNOLOGY

Communicated August 19, 1930

1. The recent experiments of Bothe and Kolhoerster[1] have led to an important and interesting conclusion. The ionization phenomena, by means of which cosmic rays are measured, are produced by highly penetrating corpuscular rays. It remains, however, an open question whether this corpuscular radiation is identical with the cosmic rays themselves, or is secondary in its nature, produced by primary rays of an electromagnetic character. Bothe and Kolhoerster admit both possibilities, but they seem to lean toward the first, owing to the fact that their corpuscular rays show a penetration of the same order of magnitude as that of the cosmic rays themselves. Further studies on the nature of these corpuscular rays are due to Rossi,[2] Tuve[3] and Mott-Smith.[4] Rossi gives some preliminary results about the deflection of these rays in a magnetic field. If they are fast-moving electrons, he finds that their energy is a little lower than 10^9 volt. He admits, however, that his observations were not sufficiently numerous to make this figure conclusive.

The purpose of the present note is to find out whether the hypothesis of an electronic nature of the primary cosmic rays is consistent with their observed intensity distribution on the surface of the earth. For lack of a better figure, we shall assume their energy to be 10^9 volt. It has been pointed out to the author by Dr. R. A. Millikan that electrons of that speed are appreciably deflected in the magnetic field of the earth. In fact, the radius of curvature a of an electron moving normally to a homogeneous magnetic field H is given by the relation

$$mv^2/a\ \sqrt{1-\beta^2} = e\beta H, \tag{1}$$

where e, m denote the charge and mass of the electron, v and $\beta = v/c$ its velocity in absolute measure and referred to the velocity of light c.

On the other hand, the energy is expressed in the following way by the voltage V

$$mc^2\left[\frac{1}{\sqrt{1-\beta^2}} - 1\right] = eV. \tag{2}$$

In the case of high speed, β is very close to 1 and $\sqrt{1-\beta^2}$ a very small number. We may neglect 1 in the bracket of Eq. (2) and, dividing the two expressions, obtain as a good approximation

$$a = V/H. \tag{3}$$

With $V = 10^9$ volt $= 0.33.10^7$ abs. and $H = 0.50$ Gauss this gives $a = 6.6.10^6$ cm. $= 66$ km. Compared with terrestrial dimensions, this is a very small radius, indeed, and this fact makes us expect a big influence of the earth field on the path of our particles.

The equations of motion turn out to be so complicated that a solution cannot readily be given in an analytical form, either rigorously or by approximations. Fortunately, however, it is possible to find certain maximum and minimum relations which give an unambiguous answer to the question we are interested in: *The assumption under discussion is inconsistent with observed facts.* Electrons of 10^9 volt energy coming from outside can strike the earth only in two limited zones around its magnetic poles. Practically all the countries where cosmic rays have been measured are outside these zones. This leaves the following possibilities: (*a*) cosmic rays are electromagnetic waves, (*b*) they are corpuscular rays of a very much higher energy than 10^9 volt (at least 6.10^{10} volt), (*c*) they are not cosmic but have terrestrial origin.

2. We treat the earth as a simple magnetic dipole and express the field in polar coordinates r, ϑ, φ

$$eH_r = -2fc\cos\vartheta/r^3,\ eH_\vartheta = -fc\sin\vartheta/r^3,\ eH_\varphi = 0. \tag{4}$$

The force acting on the electron is $\overline{F} = \frac{e}{c}\bar{v} \times \overline{H}$, with the Lagrangean components

$$\begin{cases} Q_r = f\sin^2\vartheta\dot{\varphi}/r^2,\ Q_\vartheta = -2f\sin\vartheta\cos\vartheta\dot{\varphi}/r, \\ Q_\varphi = -f\sin\vartheta\left(\dfrac{\sin\vartheta\dot{r}}{r^2} - \dfrac{2\cos\vartheta\,\dot{\vartheta}}{r}\right) \end{cases} \tag{5}$$

The kinetic potential is, therefore,

$$L = -mc^2\sqrt{1-\beta^2} + f\sin^2\vartheta\dot{\varphi}/r. \tag{6}$$

In fact, it can be easily verified that Eq. (5) are obtained by Lagrangean differentiation of the second term of this expression. The momenta, on the other hand, result by differentiation with respect to $\dot{r}$, $\dot{\vartheta}$, $\dot{\varphi}$

$$\begin{aligned} p_r &= \frac{m\dot{r}}{\sqrt{1-\beta^2}},\ p_\vartheta = \frac{mr^2\dot{\vartheta}}{\sqrt{1-\beta^2}}, \\ p_\varphi &= \frac{mr^2\sin^2\vartheta\dot{\varphi}}{\sqrt{1-\beta^2}} + \frac{f\sin^2\vartheta}{r}. \end{aligned} \tag{7}$$

and the Hamiltonian $H = -L + \dot{r}p_r + \dot{\vartheta}p_\vartheta + \dot{\varphi}p_\varphi$ assumes the form

$$H = mc^2/\sqrt{1-\beta^2}$$
$$= mc^2\left\{1+\frac{1}{m^2c^2}\left[p_\vartheta{}^2+\frac{1}{r^2}p_\varphi^2+\left(\frac{p_\varphi}{r\sin\vartheta}-\frac{f\sin\vartheta}{r^2}\right)^2\right]\right\}^{1/2}. \quad (8)$$

From this form of the Hamiltonian, there follow two facts which are obvious but which must be emphasized as they are of great importance for the following discussion. In the first place, H does not contain the time t; this leads to the existence of an equation of conservation of energy $H = mc^2 + \alpha$, where α denotes the kinetic energy.

$$mc^2/\sqrt{1-\beta^2} = mc^2 + \alpha, \quad (9)$$

or

$$p_r^2+\frac{1}{r^2}p_\vartheta^2+\left(\frac{p_\varphi}{r\sin\vartheta}-\frac{f\sin\vartheta}{r^2}\right)^2 = A^2,\ A^2 = 2m\alpha+\alpha^2/c^2. \quad (10)$$

Equation (9) *shows that the velocity remains strictly constant throughout.* This could be anticipated as the force is always normal to the path and does not produce any work.

In the second place, H does not contain the variable φ, so that $p_\varphi = p = \text{const}$. We can rewrite the third of equations (7) in the form

$$\frac{m\dot{\varphi}}{\sqrt{1-\beta^2}} = \frac{p}{r^2\sin^2\vartheta}-\frac{f}{r^3}. \quad (11)$$

When the distance of the electron from the dipole is extremely large, the term f/r^3 is negligible compared with $p/r^2\sin^2\vartheta$, and p has the meaning of an ordinary angular momentum. As the electron approaches the last term gains in importance: The angular velocity is altered by an amount proportional to r^{-3}. This is a generalization of Larmor's theorem: a homogeneous magnetic field produces the constant Larmor precession, a dipole produces a variable precession (proportional to r^{-3}) around the magnetic axis.

It follows from the two facts, just outlined, that the electron cannot come infinitely close to the dipole, except in the directions $\vartheta = 0$, or $\vartheta = \pi$. In fact, as the electron approaches the last term begins to dominate for any finite ϑ, $\dot{\varphi}$ increases as r^{-3}, and the absolute value of the azimutal component of the velocity $v_\varphi = r\sin\vartheta\dot{\varphi}$ increases as $\sin\vartheta r^{-2}$. However, we have found that the total velocity v is constant, so that $v_\varphi \leq v$ cannot increase indefinitely. This shows that, for any given ϑ, there are points which cannot be reached, or in other words: unless the electron ultimately approaches the dipole in one of the directions $\vartheta = 0$ or $\vartheta = \pi$, there is a position of nearest approach, for which $\dot{r} = p_r = 0$. Our objective will be the finding of a minimum value for this position of nearest approach.

3. With condition $p_r = 0$, Eq. (10) becomes

$$\left(\frac{p}{r\sin\vartheta} - \frac{f\sin\vartheta}{r^2}\right)^2 = B^2, \tag{12}$$

with the abbreviation

$$B^2 = A^2 - p_\vartheta^2/r^2,\ 0 \leqslant B^2 \leqslant A^2. \tag{13}$$

Solving (12) with respect to r, we find

$$\frac{1}{r} = \frac{p \pm \sqrt{p^2 \pm 4Bf\sin^3\vartheta}}{2f\sin^2\vartheta}. \tag{14}$$

This expression represents four different solutions, which we may denote by the symbols r_1 corresponding to $(+, +)$, r_2 to $(-, +)$, r_3 to $(+, -)$, r_4 to $(-, -)$. Solution r_2 must be ruled out as it is negative and has no physical significance. If we plot these solutions, for constant p, f, A, in the polar r, ϑ-plane, the solutions r_3 and r_4 represent parts of the same curve and join together in the vertex $\sin^3\vartheta = p^2/4Bf$. If we replace B by A (always $\geqslant B$)

$$\frac{1}{\rho} = \frac{p \pm \sqrt{p^2 \pm 4Af\sin^3\vartheta}}{2f\sin^2\vartheta}, \tag{15}$$

we have $\rho_1 \leqslant r_1$, $\rho_3 \geqslant r_3$, $\rho_4 \leqslant r_4$, provided that the solutions are real. This shows us, among other things, that the vertex of ρ_3, ρ_4 coincides with the vertex r_3, r_4 so that here we have $B = A$.

Let us now put the question as follows: We consider a point $r = R$, $\vartheta = \theta$ and we ask whether an electron, coming from outside, can reach it. We have to distinguish the following two cases

(I). Constant p negative, or positive and small

$$p \leqslant 2(Bf\sin^3\theta)^{1/2},$$

and *a fortiori*

$$p \leqslant 2(Af\sin^3\theta)^{1/2}.$$

This rules out solutions r_3, r_4 and ρ_3, ρ_4. The motion of the electron can take place only outside the curve ρ_1. Comparing electrons with different constants p, we see that ρ_1 is smallest for those where $p = 2(Af\sin^3\theta)^{1/2}$, namely,

$$(1 + \sqrt{2})\rho_1 = (f\sin\theta/A)^{1/2}.$$

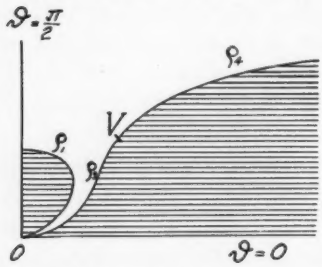
Consequently the point R, θ *cannot* be reached if $R < \rho_1$, or

$$2.41R < (f\sin\theta/A)^{1/2}. \tag{16}$$

(II). The constant p is positive and

$$p \geqslant 2(Bf \sin^3 \theta)^{1/2}.$$

In this case all three barriers r_1, r_3, r_4 of solution (14) have a physical existence, as well as the corresponding ρ_1, ρ_3, ρ_4 of (15). We give in figure 1 a graphical representation of these curves for a quadrant of the r, ϑ-plane.



The electrons, under no circumstances, can get into the shaded region. Again we have to consider two sub-cases. (IIa). R is larger than the distance of the vertex V from the origine O. This can only happen when p is small, namely, $p < 2A^2R^3/f$. With the numerical data of the next section, this turns out to be such a small number that it is permissible to neglect it altogether in the expression of the curve ρ_1 which limits the motion of our electron. R cannot become smaller than ρ_1 and this leads to the following condition.

No point can be reached for which

$$R < (f \sin \theta/A)^{1/3}. \tag{17}$$

(IIb). R is smaller than the distance of the vertex V from the origin O. In this case the electron must strike or have passed to the left side of the vertex V. We have seen that the point V is characterized by $\sin^2\vartheta = p^2/4Af$. There are points in the path of our electron for which the angle is larger than this, so that there we have $\sin^3 \vartheta \leqslant p^2/4Af$ or $p \leqslant 2(Af \sin^3 \vartheta)^{1/2}$. We can assert, *a fortiori*, $p \leqslant 2(Af)^{1/2}$. Eq. (15) with $(+, +)$ shows us that ρ has its minimum when p reaches its largest value, i.e., $2(Af)^{1/2}$. The smallest possible ρ is then

$$\rho_1 = \left(\frac{f}{A}\right)^{1/3} \frac{\sin^2 \theta}{1 + \sqrt{1 + \sin^2 \theta}}.$$

Points for which

$$R < \left(\frac{f}{A}\right)^{1/3} \frac{\sin^2 \theta}{1 + \sqrt{1 + \sin^2 \theta}} \tag{18}$$

cannot be reached.

We may leave the question open whether the curves corresponding to the different solutions (14) are simple or consist of several branches, owing to the analytical character of B. We have only made use of the fact that they lie outside the curves of Eq. (15), i.e., outside the shaded region of figure 1. Since this is assured, our results will always apply.

4. We are now ready for numerical applications to the field of the earth, which is roughly equivalent to a dipole of the strength $f = 1.24 \cdot 10^6$. For electrons of energy 10^9 volt we get $\alpha = 1.6 \cdot 10^{-3}$, $A = 5.3 \cdot 10^{-14}$. There-

fore, $\sqrt{f/A} = 4.8 \cdot 10^9$ cm. Since the radius of the earth is $R = 6.35 \cdot 10^8$, we can say $\sqrt{f/A} = 7.6\ R$.

If we replace in expressions (16), (17), (18) the sign of inequality by that of equality, they will give us directly the maximum angular distance θ_m from the magnetic pole, in which an electron can strike the surface of the earth. We obtain for the three special cases,

(I) $$7.6 \sin^{1/2}\theta_m = 2.41, \qquad (19)$$
$$\theta_m^{(1)} = 6°.$$

(IIa) $$7.6 \sin^{1/2}\theta_m = 1, \qquad (20)$$
$$\theta_m^{(2)} = 1°.$$

(IIb) $$7.6 \sin^2 \theta_m = 1 + \sqrt{1 + \sin^2 \theta_m}, \qquad (21)$$
$$\theta_m^{(3)} = 32°.$$

The net result of our considerations is, therefore, that electrons of 10^9 volt energy cannot hit the earth outside of two limited zones around the magnetic poles.[5] Our analysis does not permit us to say whether the maximum distance from the pole of 32° is actually reached.

[1] Bothe and Kolhoerster, *Zs. Physik*, **54**, 686 (1929).

[2] B. Rossi, *Rend. Acc. dei Lincei*, **2**, 478 (1930).

[3] M. A. Tuve, *Phys. Rev.*, **35**, 651 (1930).

[4] L. M. Mott-Smith, *Ibid.*, **35**, 1125 (1930).

[5] For electrons of energy 10^8 volt the maximum distance would be 17°, for 2.10^9 volt it would be 40°.

PERIODICITY IN SEQUENCES DEFINED BY LINEAR RECURRENCE RELATIONS

BY H. T. ENGSTROM*

DEPARTMENT OF MATHEMATICS, CALIFORNIA INSTITUTE OF TECHNOLOGY

Communicated August 18, 1930

A sequence of rational integers

$$u_0, u_1, u_2, \ldots \qquad (1)$$

is defined in terms of an initial set $u_0, u_1, \ldots, u_{k-1}$ by the recurrence relation

$$u_{n+k} + a_1 u_{n+k-1} + \ldots + a_k u_n = a, \; n \geqq 0, \qquad (2)$$

where $a_1, a_2, \ldots, a_k$ are given rational integers. The author examines (1) for periodicity with respect to a rational integral modulus m. Carmichael[1] has shown that (1) is periodic for $(a_k, p) = 1$ and has given periods (mod

m) for the case where the prime divisors of m are greater than k. The present note gives a period for (1) (mod m) without restriction on m. The results include those of Carmichael. The author also shows that if p divides a_k (1) is periodic after a determined number of initial terms and obtains a period.

The algebraic equation

$$F(x) = x^k + a_1x^{k-1} + \ldots + a_k = 0 \tag{3}$$

is said to be associated with (2). In considering (1) for the prime modulus p we may replace $F(x)$ by any polynomial $f(x)$ which is congruent to $F(x)$ (mod p) and of degree k with leading coefficient unity. The following lemma gives a convenient choice for $f(x)$.

LEMMA 1. We may choose $f(x) \equiv F(x) \pmod{p}$ with the following properties:

(1) $f(x)$ is irreducible and of degree k with leading coefficient unity.

(2) p does not divide the index of $f(x)$.

(3) If θ is a root of $f(x)$ and p contains precisely the αth power of a prime ideal $\mathfrak{p}$ in $K(\theta)$ then $f'(\theta)$ contains precisely $\mathfrak{p}^{\alpha-1+\rho}$, where $\rho = 1$ or 0 according as α is or is not divisible by p.

(4) $1 - \theta \not\equiv 0 \pmod{\mathfrak{p}^2}$ for any prime ideal divisor $\mathfrak{p}$ of p in $K(\theta)$.

If $\theta_1, \theta_2, \ldots, \theta_k$ are the roots of $f(x) = 0$, the general term of any sequence associated with $f(x)$ is given by

$$u_n = \frac{a}{f(1)} + \beta_1\theta_1^n + \beta_2\theta_2^n + \ldots + \beta_k\theta_k^n, \tag{4}$$

where

$$\beta_j = \frac{\gamma_j}{f^j(\theta^j)} + \frac{a\delta_j}{(1 - \theta_j)f'(\theta_j)} \tag{5}$$

and γ_j, δ_j are integers in $K(\theta_j)$.

Let

$$F(x) \equiv \phi_1(x)^{e_1}\phi_2(x)^{e_2} \ldots \phi_r(x)^{e_r} \pmod{p} \tag{6}$$

where the $\phi_i(x)$ are prime functions (mod p) whose degrees we denote by k_i. By the theorem of Dedekind (6) implies the prime ideal decomposition

$$p \equiv \mathfrak{p}_{1j}^{e_1}\mathfrak{q}_{2j}^{e_2} \ldots \mathfrak{p}^{e_r}, \; N\mathfrak{p}_{ij} = p^{k_i}, \tag{7}$$

in $K(\theta_j)$. Lemma 1 gives the power of these ideals dividing the denominators in (5). By use of the theorem of Fermat in algebraic fields we obtain the periodicity of (1) directly from (4).

We say that π is a general period (mod m) of the recurrence (2) if every sequence (1) satisfying (2) has the period π (mod m). The minimum period of any particular sequence (1) will be a divisor of π. We write

$e = \max e_i$ in (6) and l equal to the least common multiple of $p^{k_i} - 1$, $i = 1, 2, \ldots, r$. The following theorems give a general period of (2) for the case $(a_k, p) = 1$.

THEOREM 1. *If* $(a_k, p) \neq 1$, *and either* $a \equiv 0 \pmod{p}$ *or* $F(1) \not\equiv 0 \pmod{p}$, *then* (2) *has the general period* $l \pmod{p}$.

THEOREM 2. *If* $(a_k, p) = 1$ *and* $p^\epsilon \leqq e < p^{\epsilon+1}$, $\epsilon \geqq 0$, *then* (2) *has the general period* $p^{\epsilon+1}\, l \pmod{p}$.

If p divides a_k we obtain the following theorem from (4).

THEOREM 3. *If the last* s *coefficients in* (2) *are divisible by* p, $a^{k-s} \not\equiv 0 \pmod{p}$, *then* (1) *is periodic* $\pmod{p}$ *after* s *terms and a period is given by Theorems* 1 *and* 2.

A general period π of (2) for the prime power modulus p^α is obtained from a period (mod p) by the following theorem which may be proved quite directly by noting that $(u_{n+\pi} - u_n)/p$ is again a sequence satisfying the recurrence (2) with $a = 0$.

THEOREM 4. *If* $(a_k, p) = 1$, *and the recurrence* (2) *has the general period* $\pi \pmod{p}$ *then it has the general period* $p^{\alpha-1}\pi \pmod{p^\alpha}$.

If a_k is divisible by p we obtain the theorem:

THEOREM 5. *If the last* s *coefficients of* (2) *are divisible by* p, $a_{k-s} \not\equiv 0 \pmod{p}$ *then* (1) *is periodic* $\pmod{p^\alpha}$ *after* αs *terms and a period is given by Theorem* 4.

The following theorem suffices for obtaining a period of (1) for a general rational integral modulus m from the previous results.

THEOREM 6. *If* m *has the prime decomposition* $p_1^{\alpha_1} p_2^{\alpha_2} \ldots p_t^{\alpha_t}$, *the least common multiple of a set of general periods* π_i *of* (2) $(\text{mod } p_i^{\alpha_i})$, $i = 1, 2, \ldots t$ *is a general period of* (2) $(\text{mod } m)$.

* NATIONAL RESEARCH FELLOW.

[1] R. D. Carmichael, "On Sequences of Integers Defined by Recurrence Relations," *Quart. J. Math.*, **48**, 343–372 (1920).